Biometrika (2009), **96**, 2, *pp*. 399–410
doi: 10.1093/biomet/asp006
Advance Access publication 21 April 2009

Optimal testing of multiple hypotheses with common effect direction

BY RICHARD M. BITTMAN

Bittman Biostat, Inc., Glencoe, Illinois 60022, U.S.A.
rmb@bittmanbiostat.com

JOSEPH P. ROMANO

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.
romano@stanford.edu

CARLOS VALLARINO

Analytical Science, Takeda Global Research and Development, Deerfield, Illinois 60015, U.S.A.
cvallarino@tpna.com

AND MICHAEL WOLF

Institute for Empirical Research in Economics, University of Zurich, CH-8006 Zurich, Switzerland
mwolf@iew.uzh.ch

SUMMARY

We present a theoretical basis for testing related endpoints. Typically, it is known how to construct tests of the individual hypotheses, but not how to combine them into a multiple test procedure that controls the familywise error rate. Using the closure method, we emphasize the role of consonant procedures, from an interpretive as well as a theoretical viewpoint. Surprisingly, even if each intersection test has an optimality property, the overall procedure obtained by applying closure to these tests may be inadmissible. We introduce a new procedure, which is consonant and has a maximin property under the normal model. The results are then applied to PROactive, a clinical trial designed to investigate the effectiveness of a glucose-lowering drug on macrovascular outcomes among patients with type 2 diabetes.

Some key words: Closure method; Consonance; Familywise error rate; Multiple endpoints; Multiple testing; O'Brien's method.

1. INTRODUCTION

In research and experimentation it is common to specify several hypotheses. In clinical research, these have been characterized as primary, usually one hypothesis or 'endpoint', and secondary, one or more endpoints to be tested if the primary endpoint is significant. More frequently now, clinical trials feature multiple, co-primary endpoints, the significance of any of which forms the basis for a claim of efficacy. Therefore, for both scientific and regulatory reasons, the

familywise error rate for the family of co-primary endpoints must be controlled. Furthermore, it may be reasonable to expect that every co-primary endpoint will exhibit an effect of treatment, possibly some to a greater degree than others. This is the common effect direction alluded to in the title.

The desire to focus power on a common direction led O'Brien (1984) to combine multiple test statistics into a single hypothesis test. Under a normal model assumption, O'Brien derived an ordinary least-squares test statistic and a generalized least-squares test statistic that are more powerful than Hotelling's T^2 statistic in the case of related endpoints. Lehmacher et al. (1991) apply O'Brien's test in combination with the closure principle of Marcus et al. (1976). They point out that the Bonferroni test, and by extension, stepdown tests based on the maximum test statistic (Romano and Wolf, 2005), are useful for detecting one highly significant difference, or treatment effect, among a group of otherwise barely significant or nonsignificant differences. On the other hand, O'Brien's tests, based on the unweighted or weighted sum of test statistics, succeed in rejecting the global null against alternatives closer to the diagonal, by which is meant a group of similar treatment effects. Pocock et al. (1987) extend this approach to a general situation of asymptotically normal test statistics. Summing test statistics in the multivariate survival analysis setting, as we do in the example later, became theoretically justified with the method of Wei et al. (1989). The main problem we consider is how to combine tests of individual hypotheses into a multiple testing procedure that is sensitive or powerful when the endpoints are related.

Suppose that data X are available, whose distribution is given by a model $P = \{P_\theta,\ \theta \in \Omega\}$. The parameter space Ω can be parametric, semiparametric or nonparametric, since θ merely indexes the parameter space. In order to devise a procedure which controls the familywise error rate, the closure method reduces the problem to constructing tests that control the usual probability of type 1 error. To be specific, for a subset $K \subseteq \{1, \ldots, s\}$ and $\omega_i \subset \Omega$, let H_K denote the intersection hypothesis defined by

$$H_K = \omega_K \equiv \bigcap_{i \in K} \omega_i; \tag{1}$$

that is, H_K is true if and only if $\theta \in \bigcap_{i \in K} \omega_i$. Of course, $H_i = H_{\{i\}}$. Suppose that ϕ_K is an α-level test of H_K, that is, $\sup_{\theta \in \omega_K} E_\theta\{\phi_K(X)\} \leqslant \alpha$. Then the decision rule that rejects H_i if H_K is rejected for all subsets K for which $\{i\} \subseteq K$ strongly controls the familywise error rate.

Consider the choice of tests of H_K. Even in the case $s = 2$, little formal theory exists in the design of tests of H_K, but many ad hoc procedures have been developed; see Hochberg & Tamhane (1987), Westfall & Young (1993), Romano & Wolf (2005) and references therein. These approaches incorporate the dependence structure of the data and improve on Holm's (1979) method.

Stepdown tests based on the maximum test statistic yield multiple test procedures which satisfy a property called consonance; for a discussion of such tests see Remark 2 or Romano & Wolf (2005). A testing method is consonant when the rejection of an intersection hypothesis implies the rejection of at least one of its component hypotheses. An associated concept is that of coherence, which states that the nonrejection of an intersection hypothesis implies the nonrejection of any subset hypothesis it implies. Coherence is de facto true in any closed testing method. Consider a randomized experiment for testing the efficacy of a drug versus a placebo with two primary endpoints in a closed test setting: testing for reduction in headaches, H_1, and testing for reduction in muscle pain, H_2. If the joint intersection hypothesis $H_{\{1,2\}}$ is rejected but neither individual hypothesis is rejected, then one might conclude that the drug has some beneficial effect, but compelling evidence has not been established to promote a particular drug indication. Lack of

consonance, which is alternatively called dissonance, makes interpretation awkward. Moreover, we will argue that, in the framework with which we are concerned here, dissonance is undesirable in that it results in decreased ability to reject false null hypotheses.

Sonnemann & Finner (1988) showed that any incoherent procedure can be replaced by a coherent one which is at least as good. Sonnemann (1982) also showed that all coherent procedures that control the familywise error rate must be obtained by the closure method. Therefore, our restriction to procedures based on the closure method is no restriction at all. Moreover, it has been shown, in unpublished work by one of the authors, that any procedure that is not consonant can be replaced by a consonant one which is at least as good, in the sense that the familywise error rate is still controlled and there are at least as many rejections as the original procedure. This paper provides an explicit construction that yields a strict improvement over existing methods in the context of testing multiple endpoints with common effect direction.

2. Rationale for the Sum Test

In this section, we consider a stylized version of the problem. The parametric structure we now assume is an asymptotic approximation to the more general nonparametric framework. Think of X_i as denoting a test statistic for the ith hypothesis, and assume that $(X_1, \ldots, X_s)$ is multivariate normal with $X_i \sim N(\theta_i, 1)$ and known covariance matrix Σ. Let $\theta = (\theta_1, \ldots, \theta_s)$. For testing one-sided alternatives in this parametric model, the parameter space is given by

$$\Omega = \left\{ \theta : \bigcap_{i=1}^{s} \{\theta_i : \theta_i \geqslant 0\} \right\}. \tag{2}$$

However, we will also consider two-sided alternatives, but with the restriction that alternatives $(\theta_1, \ldots, \theta_s)$ are such that all θ_i have the same sign, possibly negative; that is, we will also later consider the larger parameter space

$$\Omega' = \left\{ \theta : \bigcap_{i=1}^{s} \{\theta_i : \theta_i \geqslant 0\} \right\} \bigcup \left\{ \theta : \bigcap_{i=1}^{s} \{\theta_i : \theta_i \leqslant 0\} \right\}. \tag{3}$$

For testing $H_i : \theta_i = 0$ against $\theta_i > 0$, the test that rejects H_i if $X_i > z_{1-\alpha}$ is uniformly most powerful level α. In order to apply closure, we consider tests of the intersection hypothesis $\theta_i = 0$ for all i. The general intersection hypothesis H_K given in (1) can be handled in the same way by just considering $i \in K$.

Proposition 1. *Consider the multivariate location model with mean vector $\theta \in \Omega$ and known nonsingular covariance matrix Σ, where the parameter space Ω is given by* (2). *Then*
(i) *for testing $\theta_i = 0$ for all i against the fixed alternative $(\theta_1', \ldots, \theta_s')$, the most powerful test rejects for large values of $(\theta')^{\mathrm{T}} \Sigma^{-1} X$, where X is a column vector with transpose $X^{\mathrm{T}} = (X_1, \ldots, X_s)$ and θ' is a column vector with transpose $(\theta')^{\mathrm{T}} = (\theta_1', \ldots, \theta_s')$. In particular, no uniformly most powerful test exists;*
(ii) *for testing $\theta_i = 0$ for all i against alternatives $(\theta_1', \ldots, \theta_s')$ such that all θ_i' are equal, a uniformly most powerful test exists and rejects for large values of the sum of the components of $\Sigma^{-1} X$; and*
(iii) *if, in addition, Σ has diagonal elements* 1 *and off-diagonal elements ρ, then a uniformly most powerful level α test exists and rejects the hypothesis that all $\theta_i = 0$ when $\sum_i X_i > z_{1-\alpha}\{s + s(s-1)\rho\}^{1/2}$.*

All proofs are given in the Appendix. Thus, rejecting the intersection hypothesis for large values of the sum $\sum_i X_i$ is uniformly most powerful, but only for a restricted alternative parameter space, and under a strong assumption on Σ. We now state a maximin result that applies to a much larger alternative parameter space.

PROPOSITION 2. *Assume that* Σ *has diagonal elements* 1 *and off-diagonal elements* ρ, *and consider testing* $H_0 : \theta = (0, 0, \ldots, 0)$ *against* $\theta \in \omega_1(\epsilon)$, *where*

$$\omega_1(\epsilon) = \left\{ \theta : \bigcap_i \{\theta_i : \theta_i \geqslant \epsilon\} \right\}. \tag{4}$$

Then the test that rejects when $\sum_i X_i > z_{1-\alpha}\{s + s(s-1)\rho\}^{1/2}$ *is maximin; that is, it maximizes* $\inf\{\theta \in \omega_1(\epsilon) : \mathrm{pr}_\theta(\text{reject } H_0)\}$.

Remark 1. The covariance structure of Proposition 2, known as compound symmetry, is a tractable correlation model that is used in a number of practical situations, such as repeated measures analysis of variance. Unfortunately, if Σ has a different structure, the less tractable linear combination $1'\Sigma^{-1}X$ is maximin. Note the similarity of this test statistic, derived here by testing and maximizing power, to O'Brien's (1984, p. 1082) best linear unbiased estimator of the common mean of possibly correlated random variables.

Finally, for two-sided alternatives with parameter space Ω' given in (3), an analogous maximin result holds for the test that rejects for large values of $|\sum X_i|$.

3. OPTIMAL CONSONANT TESTS

Formally, with consonant methods, if the intersection hypothesis H_K defined in (1) is rejected, then some H_i with $i \in K$ is rejected. We concentrate now on how to choose consonant tests of an intersection hypothesis. What follows is an example of a dissonant test.

Example 1. *One-sided normal means*. Recall the setup in § 2 and Proposition 2. If $\alpha = 0{\cdot}05$, $\rho = 0$ and $(X_1, X_2) = (1{\cdot}4, 1{\cdot}4)$, then no H_i can be rejected by closure, even though $H_{\{1,2\}}$ is rejected because the sum test rejects if $X_1 + X_2 > 2{\cdot}326$; see Fig. 1(a).

This procedure can be improved if the goal is to make correct decisions about H_1 and H_2. There are points in the rejection region for testing the intersection hypothesis $H_{\{1,2\}}$ that do not allow for rejection of either H_1 or H_2. By removing such points from the rejection region when testing $H_{\{1,2\}}$, we can instead include other points in the rejection region that satisfy the constraint that the overall rule be consonant, while still maintaining error control. To achieve that for our overall test of $H_{\{1,2\}}$, we restrict attention to tests that have a rejection region in the plane which lies entirely in $\{(X_1, X_2) : \max(X_1, X_2) > z_{1-\alpha}\}$. Any intersection test satisfying this constraint will result in a consonant procedure when applying the closure method.

To see a concrete way to improve upon the above procedure, consider a rejection region S_α for $H_{\{1,2\}}$ of the form

$$S_\alpha = \{(X_1, X_2) : X_1 + X_2 > s(1-\alpha), \ \max(X_i) > z_{1-\alpha}\}, \tag{5}$$

where the constant $s(1-\alpha)$ is determined so that, under $(\theta_1, \theta_2) = (0, 0)$, the region S_α has probability α. The rejection region has been obtained from Proposition 2 by removing points that do not support consonance and including points that do. The chance of rejecting any false individual null hypothesis now increases when the closure method is applied. Indeed, $s(1-\alpha) < 2^{1/2} z_{1-\alpha}$. For an illustration with $\alpha = 0{\cdot}05$, see Fig. 1(a). It follows in general that, for

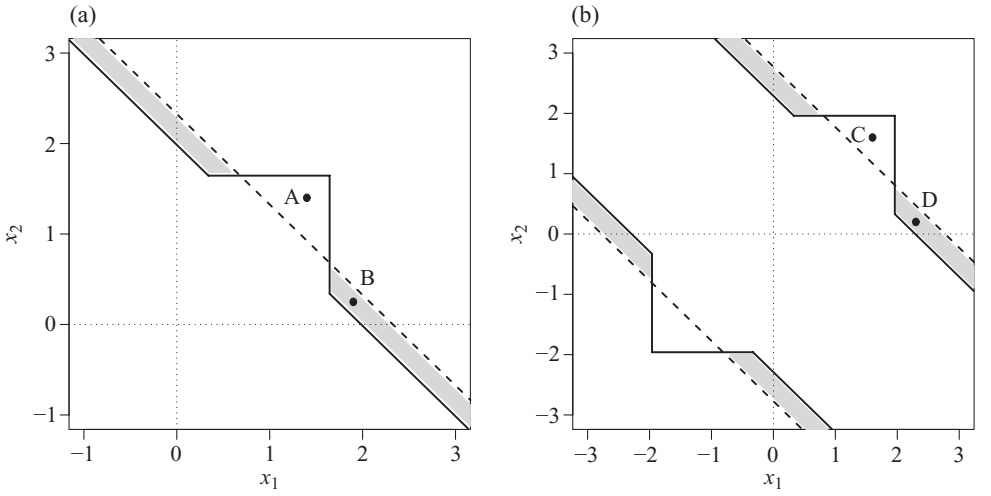

Fig. 1. (a) The rejection regions for the test of Proposition 2 and its improvement of Proposition 3 with nominal level $\alpha = 0{\cdot}05$ when the correlation is $\rho = 0$. The test of Proposition 2 rejects for points to the right and above the dashed line with intercept 2·326 and slope -1. The improved test of Proposition 3 rejects for points to the right and above the solid curve defined by (5) with $s(0{\cdot}95) = 1{\cdot}985$ and $z_{1-\alpha} = 1{\cdot}645$. For example, the point $A = (1{\cdot}4, 1{\cdot}4)$ leads to a rejection by the test of Proposition 2 but not by the improved test. On the other hand, the point $B = (1{\cdot}9, 0{\cdot}25)$ leads to a rejection by the improved test of Proposition 3 but not by the test of Proposition 2. The region which leads to at least one individual rejection by the improved test but not by the standard test is shaded. (b) The rejection regions for the test of the two-sided version of Proposition 2 based on the absolute sum, and its improvement of Proposition 4 with nominal level $\alpha = 0{\cdot}05$ when the correlation is $\rho = 0$. The absolute sum test rejects for points outside the dashed band. The improved test rejects for points outside the solid band. For example, the point $C = (1{\cdot}6, 1{\cdot}6)$ leads to a rejection by the absolute sum test (though no individual H_i are rejected), but not by the improved test of Proposition 4. On the other hand, the point $D = (2{\cdot}3, 0{\cdot}2)$ leads to a rejection by the improved test but not by the absolute sum test. The region which leads to at least one individual rejection by the improved test but not by the absolute test is shaded.

any $i = 1, 2$, with $\theta_i > 0$,

$$\mathrm{pr}_{\theta_1,\theta_2}(\text{reject } H_i \text{ using Proposition 2}) < \mathrm{pr}_{\theta_1,\theta_2}(\text{reject } H_i \text{ using } S_\alpha);$$

that is, the new consonant procedure has uniformly greater power for detecting a false null hypothesis H_i than the dissonant procedure using the sum statistic for the intersection test. Similarly, if both nulls are false, the new procedure has a uniformly greater chance of detecting both hypotheses as false or at least one false hypothesis. In summary, imposing consonance makes interpretation easier and provides better discrimination.

Thus, applying closure based on intersection tests which each have an optimality property need not result in an overall optimal procedure for the multiple testing problem. We now pursue the construction of an optimal choice of the intersection test, which will justify the use of (5). The following is a modest generalization of the Neyman–Pearson lemma, where we now impose the added consonance constraint that the rejection region be restricted to a region R of the sample space.

LEMMA 1. *Suppose that P_0 and P_1 are probability distributions with densities p_0 and p_1 with respect to a dominating measure. Restrict attention to tests $\phi = \phi(X)$ that are level α, that is, $E_0\{\phi(X)\} \leqslant \alpha$, and such that $\phi(X) = 0$ if $X \in A$, for some fixed region A in the sample space. Let $R = A^c$ be the complement of A. Among such tests, a test that maximizes the power against*

P_1 is given by

$$\phi(x) = \begin{cases} 1 & L(x) > C \quad (x \in R), \\ \gamma & L(x) = C \quad (x \in R), \\ 0 & L(x) < 0 \quad (x \in A), \end{cases}$$

where $L(x) = p_1(x)/p_0(x)$ and C and γ are chosen to meet the level constraint.

Next, we construct a maximin test by generalizing Theorem 8.1.1 in Lehmann & Romano (2005), except that now we have the added constraint that the rejection region must lie in some fixed set R. Denote by ω the null hypothesis parameter space and by ω' the alternative hypothesis parameter space over which it is desired to maximize the minimum power. The goal is to determine the test that maximizes $\inf_{\theta\in\omega'} E_\theta\{\phi(X)\}$ subject to $\sup_{\theta\in\omega} E_\theta\{\phi(X)\} \leqslant \alpha$ and to the constraint that the rejection region must lie entirely in a fixed subset R. Let $\{P_\theta, \theta \in \omega \cup \omega'\}$ be a family of probability distributions over a sample space $(\mathcal{X}, \mathcal{A})$ with densities $p_\theta = dP_\theta/d\mu$ with respect to a σ-finite measure μ, and suppose that the densities $p_\theta(x)$ considered as functions of the two variables (x, θ) are measurable with respect to $\mathcal{A} \times \mathcal{B}$ and $\mathcal{A} \times \mathcal{B}'$, where $\mathcal{B}$ and $\mathcal{B}'$ are given σ-fields over ω and ω'.

THEOREM 1. *For any distributions Λ and Λ' over $\mathcal{B}$ and $\mathcal{B}'$, for testing $h(x) = \int_\omega p_\theta(x)\, d\Lambda(\theta)$ against $h'(x) = \int_{\omega'} p_\theta(x)\, d\Lambda'(\theta)$, let $\varphi_{\Lambda,\Lambda'}$ be the most powerful among level α tests ϕ that also satisfy $\phi(x) = 0$ if $x \in R^c$. Also, let $\beta_{\Lambda,\Lambda'}$ be its power against the alternative h'. If Λ and Λ' satisfy $\sup_\omega E_\theta\{\varphi_{\Lambda,\Lambda'}(X)\} \leqslant \alpha$, $\inf_{\omega'} E_\theta\{\varphi_{\Lambda,\Lambda'}(X)\} = \beta_{\Lambda,\Lambda'}$, then $\varphi_{\Lambda,\Lambda'}$ maximizes $\inf_{\omega'} E_\theta\varphi(X)$ among all level-α tests $\phi(\cdot)$ of the hypothesis $H : \theta \in \omega$ which also satisfies $\phi(x) = 0$ if $x \in R^c$, and it is the unique test with this property if it is the unique most powerful level-α test among tests that accept on R^c for testing h against h'.*

Example 2. *Continuation of Example* 1. Recall that (X_1, X_2) is bivariate normal with unit variances, $E(X_i) = \theta_i$, and known correlation coefficient ρ. We test the null hypotheses $H_i : \theta_i = 0$ against the one-sided alternatives $\theta_i > 0$. Theorem 1 implies the following.

PROPOSITION 3. *Consider the above multiple testing problem. Apply the closure method using the test that rejects H_i if $X_i > z_{1-\alpha}$. The test of $H_{\{1,2\}}$ which maximizes*

$$\inf_{\omega_1(\epsilon)} \mathrm{pr}_{\theta_1,\theta_2}(\text{reject at least one } H_i)$$

among procedures controlling the familywise error rate is given by (5)*, where $\omega_1(\epsilon)$ is given by* (4)*.*

Remark 2. *Test based on the maximum test statistic.* In the above multiple testing problem, the test based on the maximum test statistic works as follows. Denote the ordered test statistics by $T_{(1)} \leqslant T_{(2)}$ with corresponding hypotheses $H_{(1)}$ and $H_{(2)}$. Reject $H_{(2)}$ if $T_{(2)} > q(1-\alpha)$, where the critical value $q(1-\alpha)$ depends on ρ and satisfies

$$\mathrm{pr}_{0,0}\{\max(X_1, X_2) > q(1-\alpha)\} = \alpha.$$

If $H_{(2)}$ is not rejected, stop. Otherwise, further reject $H_{(1)}$ if $T_{(1)} > z_{1-\alpha}$.

Example 3. *Application to restricted two-sided testing*. Consider the setup of § 2, except that now we consider the two-sided case. The full parameter space is given by (3) and H_i specifies $\theta_i = 0$. Here, (X_1, X_2) is bivariate normal with unit variances, $E(X_i) = \theta_i$ and known correlation

Table 1. *The critical values* $r(0{\cdot}90), r(0{\cdot}95)$ *and* $r(0{\cdot}99)$ *as functions of the correlation coefficient* ρ. *These values were obtained from* $B = 10^6$ *Monte Carlo simulations*

ρ	$r(0{\cdot}90)$	$r(0{\cdot}95)$	$r(0{\cdot}99)$	ρ	$r(0{\cdot}90)$	$r(0{\cdot}95)$	$r(0{\cdot}99)$
				0·0	1·982	2·290	2·878
−0·1	1·804	2·075	2·596	0·1	2·152	2·498	3·153
−0·2	1·617	1·853	2·313	0·2	2·314	2·700	3·421
−0·3	1·423	1·622	2·031	0·3	2·467	2·892	3·687
−0·4	1·228	1·392	1·743	0·4	2·611	3·071	3·943
−0·5	1·024	1·160	1·452	0·5	2·746	3·240	4·194
−0·6	0·819	0·930	1·161	0·6	2·872	3·401	4·427
−0·7	0·614	0·696	0·866	0·7	2·988	3·548	4·634
−0·8	0·409	0·465	0·581	0·8	3·095	3·683	4·827
−0·9	0·205	0·231	0·291	0·9	3·197	3·807	5·003
−1·0	0·000	0·000	0·000	1·0	3·290	3·920	5·152

coefficient ρ. We now determine the consonant, maximin, level-α test against $\omega_1'(\epsilon)$ defined by

$$\omega_1'(\epsilon) = \{\theta : \theta_i \geqslant \epsilon, i = 1, 2\} \cup \{\theta : \theta_i \leqslant -\epsilon, i = 1, 2\}. \tag{6}$$

PROPOSITION 4. *Consider the above multiple testing problem. Apply the closure method using the test that rejects* H_i *if* $|X_i| > z_{1-\alpha/2}$. *The test of* $H_{\{1,2\}}$ *which maximizes*

$$\inf_{\omega_1'(\epsilon)} \mathrm{pr}_{\theta_1,\theta_2}(\text{reject at least one } H_i)$$

among procedures controlling the familywise error rate is given by

$$\{(X_1, X_2) : |X_1 + X_2| > r(1-\alpha), \ \max(|X_i|) > z_{1-\alpha/2}\},$$

where $r(1-\alpha)$ *is determined so that the region has probability* α *under* $(\theta_1, \theta_2) = (0, 0)$.

Again, the optimal region takes the same form as the one without restricting to consonant tests, but adds the necessary restriction on the rejection region. For an illustration, see Fig. 1(b). Table 1 shows the critical values $r(0{\cdot}90), r(0{\cdot}95)$ and $r(0{\cdot}99)$ as functions of ρ. By symmetry, $s(1-\alpha) = r(1-2\alpha)$, so some one-sided critical values can also be derived from the table. The critical values $r(1-\alpha)$ in Table 1 were obtained by simulation. To see how, fix α. Draw B random samples from the bivariate normal distribution with means 0, unit variances and correlation coefficient ρ. Call the bth such sample $\{X_1^*(b), X_2^*(b)\}$. If $\max_{i\in\{1,2\}}\{|X_i^*(b)|\} > z_{1-\alpha/2}$, let $Y(b) = |\sum_{i=1}^2 X_i^*(b)|$; otherwise, let $Y(b) = 0$. Then $r(1-\alpha)$ is obtained as the empirical $1-\alpha$ quantile of the B values $Y(1), \ldots, Y(B)$.

Table 2 compares $\hat{\alpha}$, the empirical familywise error rate, and $\hat{\beta}$, the empirical $\mathrm{pr}_{\theta_1,\theta_2}(\text{reject} \geqslant 1 \text{ false } H_i)$, of the Holm and stepwise maximum test statistic tests to the standard and consonant sum tests. Remark 2 provides a definition of the stepwise maximum test statistic test in this context; the corresponding critical values $q(1-\alpha)$ were also obtained by simulation. Each scenario is based on 50 000 simulations from a bivariate normal distribution with mean vector (θ_1, θ_2), unit variances and correlation coefficient $\rho = 0$ and 0·5, with one-sided $\alpha = 0{\cdot}025$. The maximum test statistic is more powerful than the Holm one for nonzero ρ, and the consonant sum test is always more powerful than the sum test. As expected, the maximum test statistic is most powerful when there is only one nonzero mean, while the consonant sum test is most powerful when there are two equal nonzero means. With two unequal nonzero means, one large and one medium-sized, the consonant sum test is more powerful for $\rho = 0$, while the maximum test statistic is more

Table 2. *Comparison of empirical familywise error rate, denoted by $\hat{\alpha}$, and empirical power, denoted by $\hat{\beta}$, attained by different methods as functions of the correlation ρ and the true means (θ_1, θ_2)*

ρ	(θ_1, θ_2)	Holm	maxT	Sum	ConS	Holm	maxT	Sum	ConS
			$\hat{\alpha}$				$\hat{\beta}$		
0·0	(0·0, 0·0)	0·0249	0·0251	0·0159	0·0242	0	0	0	0
0·0	(3·0, 0·0)	0·0215	0·0215	0·0241	0·0242	0·778	0·779	0·549	0·660
0·0	(3·0, 1·5)	0	0	0	0	0·829	0·829	0·852	0·881
0·0	(3·0, 3·0)	0	0	0	0	0·951	0·951	0·976	0·978
0·5	(0·0, 0·0)	0·0235	0·0250	0·0218	0·0246	0	0	0	0
0·5	(3·0, 0·0)	0·0237	0·0237	0·0240	0·0240	0·778	0·787	0·411	0·446
0·5	(3·0, 1·5)	0	0	0	0	0·793	0·801	0·735	0·758
0·5	(3·0, 3·0)	0	0	0	0	0·898	0·904	0·925	0·931

Holm, classical Holm method; maxT, stepwise maximum test statistic; Sum, closure method using the sum statistic; ConS, consonant version of the sum test.

powerful for $\rho = 0{\cdot}5$. Due to its lack of consonance, the sum test's empirical familywise error rate sometimes falls quite short of 0·025.

Remark 3. *Control of directional errors.* Suppose that, if H_i is rejected by a given procedure, such as that of Proposition 4, then we declare $\theta_i > 0$ if $X_i > 0$ or $\theta_i < 0$ if $X_i < 0$. A directional error occurs if it is declared that $\theta_i < 0$ when in fact $\theta_i > 0$, or if it is declared that $\theta_i > 0$ when $\theta_i < 0$. Control of the familywise error rate, i.e., Type 1 errors, and directional errors together entails showing that, for any (θ_1, θ_2),

$$\mathrm{pr}_{\theta_1,\theta_2}(\text{either reject at least one true } H_i \text{ or make one or more directional errors}) \leqslant \alpha. \quad (7)$$

Application of the closure method need not result in control of directional errors; see Shaffer (1980). For some recent literature on directional errors, see Finner (1999) and Shaffer (2002). In general, the value on the left-hand side of (7) will be no smaller than the probability of at least one false rejection, the familywise error rate. Simulations over a wide range of (θ_1, θ_2) and ρ support the validity of (7) for our procedure. However, we can only argue that the procedure based on Proposition 4 satisfies (7) if α is replaced by $3\alpha/2$. If both H_i are true, there is nothing to prove, since (7) is then covered by familywise error rate control. Next, suppose that both H_i are false. If both θ_i are less than 0, then the left-hand side of (7) does not exceed

$$\mathrm{pr}_{\theta_1,\theta_2}(\text{at least one } X_i > z_{1-\alpha/2}),$$

which by Bonferroni's inequality is no bigger than

$$\mathrm{pr}_{\theta_1}(X_1 > z_{1-\alpha/2}) + \mathrm{pr}_{\theta_2}(X_2 > z_{1-\alpha/2}).$$

However, each term is bounded above by the same expression with θ_i replaced by zero, since $\theta_i < 0$, leading to the upper bound α. A similar argument holds if one θ_i is positive and the other negative, or if both are positive. The final case occurs if one H_i is true and the other false. Assume without loss of generality that $\theta_1 = 0$ and $\theta_2 < 0$. Then the event that H_1 is rejected or θ_2 is declared positive implies either $|X_1| > z_{1-\alpha/2}$ or $X_2 > z_{1-\alpha/2}$. By a Bonferroni argument, the probability of the union of these events under $(0, \theta_2)$ is bounded above by

$$\mathrm{pr}_{0,\theta_2}(|X_1| > z_{1-\alpha/2}) + \mathrm{pr}_{0,\theta_2}(X_2 > z_{1-\alpha/2}).$$

The first term equals α, and the second term is bounded above by the same expression with θ_2 replaced by 0, yielding $\alpha/2$. The sum $3\alpha/2$ is the bound.

Remark 4. *General s*. The previous results generalize to s hypotheses. For example, consider Example 3, but for general s. Let $(X_1, \ldots, X_s)$ be multivariate normal with known covariance matrix Σ and mean vector $(\theta_1, \ldots, \theta_s)$. The parameter space consists of Ω' given by (3). Assume that Σ has all off-diagonal elements equal to ρ and diagonal elements equal to one. The test that rejects for large values of $|\sum X_i|$ is maximin, but, if used for testing the intersection hypothesis that all $\theta_i = 0$, application of the closure method does not result in a consonant procedure. To see how closure leads to an improved multiple testing procedure, first test individual hypotheses H_i by rejecting H_i if $|X_i| > z_{1-\alpha/2}$. For testing the general intersection hypothesis H_K, which specifies $\theta_i = 0$ for $i \in K$, consider the following test with rejection region:

$$R_{\alpha,K} \equiv \left\{ (X_1, \ldots, X_s) : \left| \sum_{i \in K} X_i \right| > r(1-\alpha, K), \text{ and at least one } H_i, i \in K, \right.$$
$$\left. \text{is rejected when applying closure to the family } \{H_i, i \in K\} \right\},$$

where the critical value $r(1-\alpha, K)$ is determined so that the above region has probability at most α when $\theta_i = 0$ for all i. Evidently, the critical values $r(1-\alpha, K)$ must be determined inductively, so that, in order to determine $r(1-\alpha, K)$, we first determine $r(1-\alpha, K')$ for all $K' \subset K$. The test H_K is maximin among level α tests which satisfy the consonant constraint that the rejection region $R_{\alpha,K}$ must lie in $\bigcup_{K' \subset K} R_{\alpha,K'}$. Critical values may be approximated by simulation similar to the case $s = 2$. Note that $r(1-\alpha, K)$ does not depend on s and depends on K only through $|K|$.

4. Application to the PROactive Clinical Trial

To illustrate the concepts developed here, we use data from PROactive, which stands for PROspective pioglitAzone Clinical Trial In macroVascular Events, a randomized, double-blind clinical trial designed to investigate prospectively the effect of an oral glucose-lowering drug on macrovascular outcomes (Dormandy et al., 2005). The study enrolled 5238 patients with type 2 diabetes and evidence of macrovascular disease from 19 European countries. Patients were randomly assigned to either pioglitazone treatment or a placebo and were allowed to remain on whatever other anti-diabetic medication they were taking at the start of the study, except for other agents in pioglitazone's class, as well as specific cardiovascular and lipid-altering medications. The PROactive study aimed to achieve significance in a primary composite endpoint, the time to first occurrence of any of seven events: death, non-fatal myocardial infarction, including silent myocardial infarction, stroke, major leg amputation, acute coronary syndrome, leg revascularization and cardiac intervention, including coronary artery bypass graft or percutaneous coronary intervention. A second endpoint was also of interest, and consisted of a subset of the primary events: time to first occurrence among death, non-fatal myocardial infarction, excluding silent myocardial infarction, and stroke. More information on the PROactive trial can be found on the website www.proactive-results.com/index.htm.

Two interim analyses were performed using an alpha spending function, which reduced the nominal familywise error rate available at the end of the study to 0·044 from the original 0·05. After completion of the three-year study, the log-rank test (Lachin, 2000) of the primary endpoint yielded a p-value of 0·095. The log-rank test of the principal secondary endpoint yielded a corresponding p-value of 0·027. While Dormandy et al. (2005) claimed a significant outcome, critics such as Freemantle (2005) countered that a secondary endpoint cannot be deemed significant in the absence of a significant outcome in the primary endpoint, an assertion supported by Chi (1998), among others. However, if we viewed both endpoints as corresponding to hypotheses of equal interest, rather than tiered as primary and secondary, that is, if we

consider the endpoints to be co-primary, significance of the second, rather than secondary, endpoint could be scientifically assessed through a testing strategy that controls the familywise error rate. At the time of study design, the primary endpoint was defined and recognized as clinically relevant, under the assumption that all vascular beds would be equally affected by the disease state. However, the clinical relevance of the secondary endpoint was also apparent. Would the second endpoint have attained significance had the testing methods set forth in this paper been applied?

The closed family of tests for this example consists of the tests of hypotheses H_1 and H_2 of the respective co-primary endpoints and the global null hypothesis $H_{\{1,2\}}$ that neither endpoint exhibits a treatment effect. Results from tests of H_1 and H_2 are already available from the log-rank tests, as outlined above. Thus, we proceed to test the intersection hypothesis $H_{\{1,2\}}$ by applying the methods of this paper. If our test produces a p-value less than or equal to 0·044, the second endpoint could have been declared significant, not by ignoring the multiplicity problem, but by proper control of the familywise error rate.

The now-co-primary endpoints are clearly related and highly correlated, sharing the common components of death, non-fatal myocardial infarction and stroke. Even before results were published, we would have expected a treatment effect, if there were one, to be apparent in both endpoints. If we want a test of $H_{\{1,2\}}$ that directs power in the direction of our alternative hypothesis of a common effect direction, in the region $\omega_1'(\epsilon)$ of (6), the absolute sum test or its modified consonant maximin sum test of Proposition 4 are the obvious choices.

Since the log-rank test statistics are available, our first inclination might be to sum them. However, in the case of no ties, Cox (1972) derived the log-rank test as an efficient score test in a proportional hazards regression model with a single binary covariate for treatment group; see Lachin (2000). This equivalence with the proportional hazards model, which holds approximately if there are relatively few ties, allows us instead to sum the studentized parameter estimates in a simple fit of the Wei et al. (1989) marginal model with two endpoints, a relatively simple task in SAS. This sum can represent an overall treatment effect, as measured by the proportional hazards model, and corresponds to the sum of the logs of the hazard ratios. Wei et al. (1989) showed that these estimators, based on the endpoint-specific partial likelihoods, are approximately normal for large sample sizes.

Let η be the vector of $s = 2$ parameters in the Wei et al. (1989) marginal model. From the robust covariance matrix (Liang & Zeger, 1986) output by SAS, we estimate the standard errors as well as the correlation between the parameter estimates, $\hat{\rho} = 0{\cdot}74$. Studentizing the parameter estimates as $X_i = \hat{\eta}_i / \mathrm{SE}(\hat{\eta}_i)$ yields $X_1 = -1{\cdot}667$ and $X_2 = -2{\cdot}202$. We test the intersection hypothesis $H_{\{1,2\}} : \eta_1 = \eta_2 = 0$ by forming, as in Proposition 2, the test statistic $(X_1 + X_2)/(2 + 2\hat{\rho})^{1/2} = -2{\cdot}073$. The probability of a larger absolute value under the standard normal is 0·038. Since it is below 0·044, the available α, we reject the intersection hypothesis and, by the closure principle, claim that the second endpoint, indeed, had a significant treatment effect, even after accounting for multiple testing.

What result would be obtained from the PROactive data if we applied the consonant maximin sum test of Proposition 4? To calculate the critical value for this test, we drew $B = 50\,000$ random samples from the bivariate normal (0, 0, 1, 1) distribution with correlation coefficient 0·74, the observed correlation between the Wei et al. (1989) parameter estimates. Following the approach of Example 3, we generated the approximate quantile $r(1 - 0{\cdot}044) = 3{\cdot}700$; a linearly interpolated value from Table 1, at $\rho = 0{\cdot}74$, is roughly 3·768, somewhat far from the value generated through simulation, as these critical values are quite nonlinear in the inputs, especially in the level. The sum of the studentized parameter estimates has absolute value 3·869, corresponding to a p-value

of 0·036. Hence, we can again reject $H_{\{1,2\}}$ and by consonance claim significance of the second endpoint.

Acknowledgement

The authors gratefully acknowledge a referee, an associate editor and Professor D. M. Titterington for their helpful comments and suggestions that considerably improved the paper. The second author was partially supported by the U.S. National Science Foundation and the fourth author by the Spanish Ministry of Science and Technology and European Fund for Regional Development (FEDER).

Appendix

Proof of Proposition 1. The proof is an application of the Neyman–Pearson lemma. □

Proof of Proposition 2. The least favourable distribution concentrates on the single point $(\epsilon, \ldots, \epsilon)$. Maximinity results because the resulting test against this fixed alternative has an increasing power function in each of the components θ_i, and therefore the power is minimized over $\omega_1(\epsilon)$ at $(\epsilon, \ldots, \epsilon)$. □

Proof of Lemma 1. We maximize $E_{P_1}\{\phi(X)I(X \in R)\}$ subject to $E_{P_0}\{\phi(X)I(X \in R)\} \leqslant \alpha$. Let Q_i denote the conditional distribution of X given $X \in R$ when $X \sim P_i$. Also, let $\beta_i = P_i(R)$. Then, equivalently, the problem is to maximize ${\beta_1}^{-1}E_{Q_1}\{\phi(X)\}$ subject to $E_{Q_0}\{\phi(X)\} \leqslant \alpha/\beta_0$, or equivalently maximize $E_{Q_1}\{\phi(X)\}$ subject to $E_{Q_0}\{\phi(X)\} \leqslant \alpha' = \alpha/\beta_0$. By the Neyman–Pearson lemma, the optimal test rejects for large values of the likelihood ratio $dQ_1(X)/dQ_0(X)$, which is a constant multiple of $L(X)$. □

Proof of Theorem 1. If φ^* is any other level-α test of H satisfying $\varphi^*(X) = 0$ if $X \in R^c$, it is also of level α for testing the simple hypothesis that the density of X is h; therefore, the power of φ^* against h' cannot exceed $\beta_{\Lambda,\Lambda'}$. It follows that

$$\inf_{\omega'} E_\theta\{\varphi^*(X)\} \leqslant \int_{\omega'} E_\theta\{\varphi^*(X)\}\, d\Lambda'(\theta) \leqslant \beta_{\Lambda,\Lambda'} = \inf_{\omega'} E_\theta\{\varphi_{\Lambda\Lambda'}(X)\},$$

and the second inequality is strict if $\varphi_{\Lambda\Lambda'}$ is unique. □

Proof of Proposition 3. Consider any other method based on closure which rejects H_i if $X_i > z_{1-\alpha}$ and controls the familywise error rate. Let S' denote its rejection region for $H_{\{1,2\}}$. Furthermore, let $S'' = S' \cap R$, where $R = \{(X_1, X_2) : \min(X_i) > z_{1-\alpha}\}$. Then, using S' or S'' results in the same outcomes for the individual tests of H_i as far as the closure method is concerned. In particular, the probability of rejecting at least one H_i using S', combined with closure, is the same as the probability of rejecting at least one H_i using S''. However, for $S'' \subset R$, the procedure is now consonant and the probability of rejecting at least one H_i is the same as the probability of just rejecting the intersection hypothesis based on S''. The optimal choice for S'' is therefore given by Theorem 1. To apply the theorem, take Λ' to be concentrated on (ϵ, ϵ) and apply Lemma 1. The resulting test with rejection region S_α is then easily seen to be the consonant sum test given by (5). □

Proof of Proposition 4. The proof is analogous to the proof of Proposition 3, except that Λ' puts equal mass at (ϵ, ϵ) and $(-\epsilon, -\epsilon)$. □

References

Chi, G. Y. H. (1998). Multiple testings: multiple comparisons and multiple endpoints. *Drug Info. J.* **32**, 1347S–62S.

Cox, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Dormandy, J. A., Charbonnel, B., Eckland, D. A., Erdmann, E., Massi-Benedetti, M., Moules, I. K., Skene, A. M., Tan, M. H., Lefèbvre, P. J., Murray, G. D., Standl, E., Wilcox, R. G., Wilhelmsen, L., Betteridge, J.,

BIRKELAND, K., GOLAY, A., HEINE, R. J., KORÁNYI, L., LAAKSO, M., MOKÁN, M., ET AL. (2005). Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. *Lancet* **366**, 1279–89.

FINNER, H. (1999). Stepwise multiple test procedures and control of directional errors. *Ann. Statist.* **27**, 274–89.

FREEMANTLE, N. (2005). How well does the evidence on pioglitazone back up researchers' claims for a reduction in macrovascular events? *Br. Med. J.* **331**, 836–8.

HOCHBERG, Y. & TAMHANE, A. (1987). *Multiple Comparison Procedures*. New York: John Wiley.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

LACHIN, J. M. (2000). *Biostatistical Methods: The Assessment of Relative Risks*. New York: John Wiley.

LEHMACHER, W., WASSMER, G. & REITMEIR, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47**, 511–21.

LEHMANN, E. L. & ROMANO, J. P. (2005). *Testing Statistical Hypotheses,* 3rd ed. New York: Springer.

LIANG, K. Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

MARCUS, R., PERITZ, E. & GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–60.

O'BRIEN, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–87. Correction (1995), **51**, 1580–1.

POCOCK, S. J., GELLER, N. L. & TSIATIS, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–98.

ROMANO, J. P., & WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Statist. Assoc.* **100**, 94–108.

SHAFFER, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *Ann. Statist.* **8**, 1342–7.

SHAFFER, J. P. (2002). Optimality results in multiple hypothesis testing. In *The First Erich L. Lehmann Symposium – Optimality*, Ed. J. Rojo and V. Pérez-Abren. IMS Lecture Notes **44**, 11–36. Beachwood, Ohio: Inst. Math. Statist.

SONNEMANN, E. (1982). Allgemeine Lösungen multipler Testprobleme. *EDV Medizin Biol.* **13**, 120–8.

SONNEMANN, E. & FINNER, H. (1988). Vollständigkeitssätze für multiple Testprobleme. In *Multiple Hypothesenprüfung*, Ed. P. Bauer, G. Hommel and E. Sonnemann, pp. 121–35. Berlin: Springer.

WEI, L. J., LIN, D. Y. & WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Statist. Assoc.* **84**, 1065–73.

WESTFALL, P. & YOUNG, S. (1993). *Resampling-Based Multiple Testing.* New York: John Wiley.

[*Received February* 2007. *Revised August* 2008]